Dasymetric Modelling of Small-Area Population Distribution Using Land Cover and Light Emissions Data

Article *in* Remote Sensing of Environment · June 2007

DOI: 10.1016/j.rse.2006.11.020

4 authors, including:

David John Briggs
Imperial College London
180 PUBLICATIONS **5,240** CITATIONS
SEE PROFILE

Daniela Fecht
Imperial College London
33 PUBLICATIONS **478** CITATIONS
SEE PROFILE

Available online at www.sciencedirect.com
ScienceDirect
Remote Sensing of Environment 108 (2007) 451–466
Remote Sensing of Environment
www.elsevier.com/locate/rse

Dasymetric modelling of small-area population distribution using land cover and light emissions data

David J. Briggs *, John Gulliver, Daniela Fecht, Danielle M. Vienneau

Imperial College London, London, UK

Received 5 August 2006; received in revised form 26 November 2006; accepted 26 November 2006

Abstract

Despite the improvements made in census procedures over recent decades, the availability of detailed population data is limited. For many applications, including environmental and health analyses, methods are therefore needed to model population distribution at the small-area level. With the development of GIS and remote sensing techniques, the ability to develop such models has greatly improved. This paper describes a GIS-based approach using remotely sensed land cover and nighttime light emissions data to model population distribution at the land parcel level across the European Union. Light emission data from the DMSP satellites were first resampled and modelled using kriging and inverse distance weighting methods to provide a 200-m resolution light emissions map. This was then matched to CORINE land cover classes across the EU. Regression methods were used to derive models of relationships between census population counts (at NUTS 5 level) and land cover area and light emissions. Models were developed at both national and EU scale, using a range of different modelling strategies. Model performance, as indicated by the regression statistics, was seen to be good, with R^2 typically in the order of 0.8–0.9 and SEE ca. 4000 people. In southern countries, especially, incorporation of light emissions data was found to improve model performance considerably compared to models based only on land cover data. More detailed *post hoc* validation in Great Britain, using independent data on population at census tract (enumeration district and output area) and postcode level, for 1991 and 2001, showed that models gave good predictions of population at the 1 km level ($R^2 > 0.9$), but were less reliable at resolutions below ca. 500 m. Impending enhancements in the available land cover and light emissions data are expected to improve the capability of this modelling approach in the future.

Keywords: Population; GIS; Light emissions; Land cover; Spatial modelling

1. Introduction

Reliable information on population distribution is essential for a wide range of applications in both the science and policy domains. Without an adequate knowledge of where people live and spend their time, it is all but impossible to model human activities, to plan service provision, to estimate pressures on the environment, or to assess human exposures and risks to health outcome. Across most of the world, census data provide routinely available information on population numbers and composition, at least on a decennial basis. For many applications, however, these data suffer from a number of important limitations. Apart from uncertainties or biases in the censuses themselves (Boyle & Dorling, 2004; Cook, 2004), these relate mainly to the ways in which the census tracts used for collecting and reporting the data are defined. Often, these are relatively large and of variable population number, shape and size. Marked differences in both the visual representation of population distribution, and in the scale or reliability of any derived measures for which they are used as denominators (e.g., disease rates), may thus occur across a study area. Use of different spatial units may greatly change the apparent spatial patterns and associations — the classic Modifiable Areal Unit Problem, or MAUP (Openshaw & Taylor, 1981; Openshaw, 1984). Census districts change over time as a result of administrative restructuring, making analysis of long-term trends difficult. In many cases, also, census tracts do not conform to, or nest within, the other spatial structures (zone systems) for which information is available, so that population data may need to be translated between different spatial structures for the purpose of data linkage and analysis.

* Corresponding author.
E-mail address: d.briggs@imperial.ac.uk (D.J. Briggs).

0034-4257/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2006.11.020

For all these reasons, there is a need for methods to estimate small-area population numbers. Various methods have been applied for this purpose. Probably the most widely used, and simplest, is by area-weighting — that is, to redistribute populations according to the proportion of each census tract that falls within the target zones of interest (Goodchild & Lam, 1980). This is the approach used, for example, to construct the gridded population of the world (GPW-3) database (CIESIN, 2000). It is clearly highly approximate, for it relies on the (usually false) assumption that the population is evenly distributed within each census district. Another, albeit rarely used, approach is to model population patterns as continuous surfaces, for example using smoothing algorithms to fit a surface through tract centroids (Tobler, 1979; Tobler et al., 1995; Martin, 1989). Again, however, this fails to recognise that, in most areas, population patterns are highly disjunct, with large settlement centres separated by wider areas of dispersed homesteads or smaller population nuclei.

More reliable models of population distribution thus require more sophisticated techniques, and above all the use of additional data on exogenous variables that can be assumed to reflect lower level (i.e., within census tract) variations in population density. This approach has been labelled dasymetric mapping (Flowerdew & Green, 1994), and has been pursued, especially, with the help of land cover data derived from satellite imagery (Harris & Longley, 2000; Mennis, 2003).

Land cover clearly provides a useful indicator of where people live. As a basis for detailed population mapping, however, it still suffers from a major limitation — namely, how to derive weights for each land cover class or parcel that reflects its population density. This paper explores the use of nighttime light emission data, linked to land cover data, for this purpose. Models are developed and used to derive a high-resolution (200 m) population map of the European Union (EU-15, excluding Sweden and northern Finland). The work described here was undertaken as part of two EU-funded studies: MANTLE (Mapping Night-Time Light Emissions) and APMoSPHERE (Air Pollution Modelling for Support to Policy on Health and Environmental Risks in Europe). The authors gratefully acknowledge the financial support received through these two studies, and for the collaboration and assistance from the many other researchers involved.

2. Methods

2.1. Modelling strategy

Mapping population distributions on the basis of land cover class is likely to be a significant improvement on simple area-weighting methods. It is nevertheless likely to under-estimate local variations in population density for no allowance is made for intra-class (e.g., between land parcel) heterogeneity. More realistic modelling requires that weights or functions be derived to represent these intra-class variations. In principle, these may be derived in a variety of ways. They may, for example, be imputed *a priori* (on the basis of pre-existing knowledge). They may be derived stochastically by analysing relationships between land cover distribution and population numbers. They may be based on known covariates of population distribution derived either from exogenous sources or secondary features of the land cover classes: Chen (2002), for example, proposed using textural information (derived for small windows — 5×5 or 7×7 pixels) to help assess housing density in areas classified as residential land. In some cases, also, independent information on population distribution may be available, in the form of cadastral, postcode or address point data.

Whatever approach is used, the models developed must satisfy four crucial criteria:

1. Population estimates must be provided for every land parcel;
2. All population estimates should be non-negative;
3. Population estimates should be non-stationary (i.e., should be free to vary from one land parcel to another);
4. Errors in population estimates should be intrinsic to each census tract; thus models should be pycnophylactic (Tobler, 1979; Flowerdew and Green, 1994), such that the populations derived sum to actual totals for larger, containing census tracts, for which the populations are known.

The approach developed here uses data on nighttime light emissions as covariates of population density. Potential advantages of these data are that they: (a) may be expected to provide good proxies for population distribution; (b) are readily available on a world-wide basis; and (c) are updated regularly, so that they might be used to model short-term changes in population.

The light emissions data used are derived from the Defense Meteorological Satellite Program (DMSP), operated by the US Air Force. This programme first provided low-light imaging data in 1972, and now provides a routine source of nighttime light emissions data. The DMSP satellites are in low-altitude (830 km) sun-synchronous polar orbits, and carry oscillating scan radiometers – Operational Linescan Systems (OLS) – with low-light visible and thermal infrared imaging capabilities. Whilst the main purpose of these satellites is to monitor global weather conditions during daylight hours, at least one sensor has been regularly operated at night, at a gain setting capable of detecting clouds using moonlight. To achieve this, the sensor on the OLS intensifies the observed VNIR radiance in the 0.5 to 0.9 μm waveband using a photomultiplier tube. As a consequence, faint sources of VNIR emission can be detected on the Earth's surface, and the sensor is reported to be four orders of magnitude more sensitive than other, currently available, satellite sensors. Under cloudless conditions, they thus yield data on nighttime light emissions from ground-level sources.

The use of light emissions data as a proxy of population distribution and density has received growing attention in recent years. Data from DMSP have already been used, for example, to map patterns of human settlement at both continental and national or regional level (Nizeyimana et al., 2001; Pozzi et al., 2002). Broad scale relationships between nighttime light emissions and population density have likewise been demonstrated across a number of countries and used as a basis for mapping population distribution (Elvidge et al., 1997a,b;

Imhoff et al., 1997; Sutton et al., 2001). Changes in light emissions over time have also been used as an indication of urban development both at the national level and for individual cities (Cinzano, 2000; Lawrence et al., 2002).

Use of nighttime light emission data as covariates for population density nevertheless faces a number of problems. One of these is the spatial resolution of the available data. In the past, data were collected and provided at two resolutions: a fine resolution of ca. 500 m (ground sample distance), based on the original observed data, and coarse resolution of 2700 m, derived from on-board averaging of 5×5 pixels. At the time of this study, technical problems with the on-board storage devices meant that only the coarser resolution data were available. The resolution of these data inhibits their use for small-area population mapping. The second difficulty is caused by 'blooming' due to surface reflection and scattering and refraction in the atmosphere. This, too, generates ambiguity in the signal received by the sensor, and degrades the image to some extent. The third problem relates to temporal variations in light emissions, as a result of changes in source activity, weather conditions (and thus variations in surface reflection and atmospheric scattering), cloud cover and natural light (especially moonlight). For this reason, data from a number of satellite images usually need to be averaged in order to estimate the so-called stable light emissions. The fourth problem relates to likely variations in the relationship between light emissions and population density from one area to another. This is due in part to contributions from a wide range of non-residential sources, including transport, recreational, commercial and industrial activities. It is also due to variations in affluence

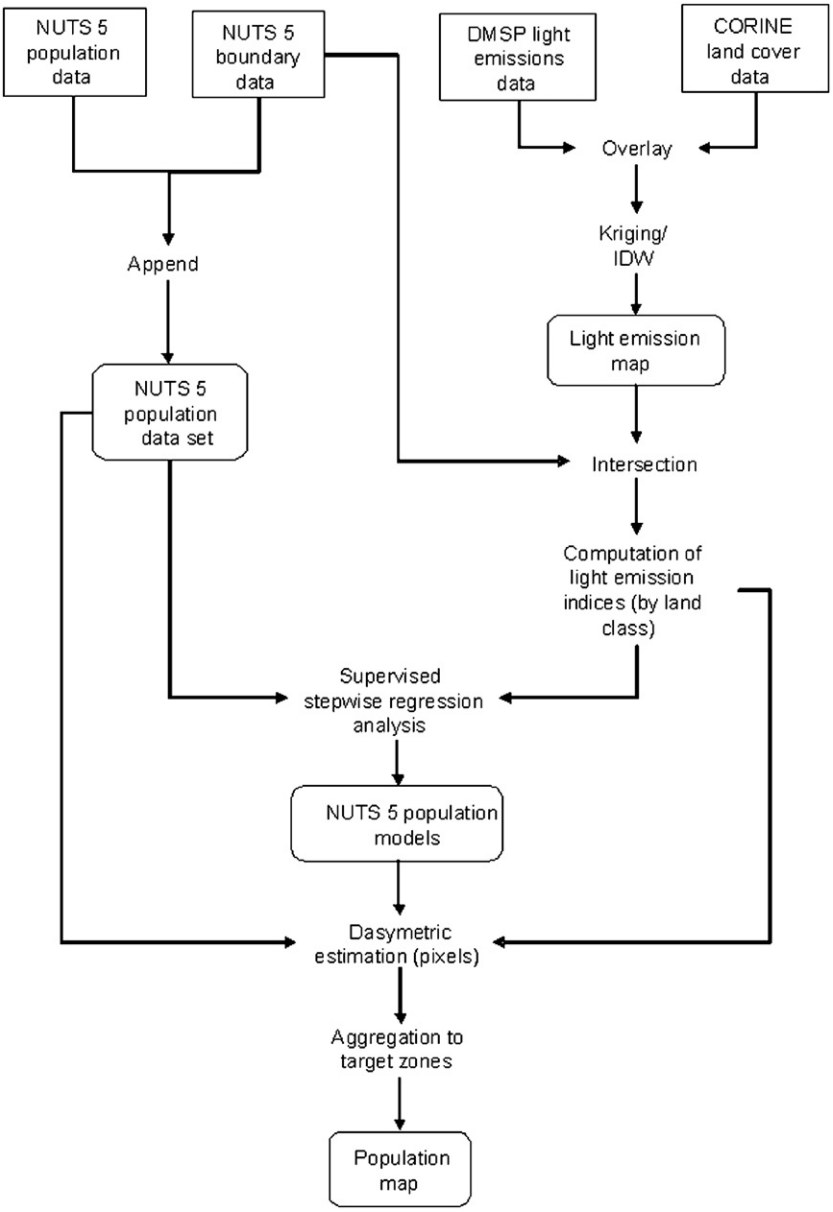

Fig. 1. Population modelling strategy.

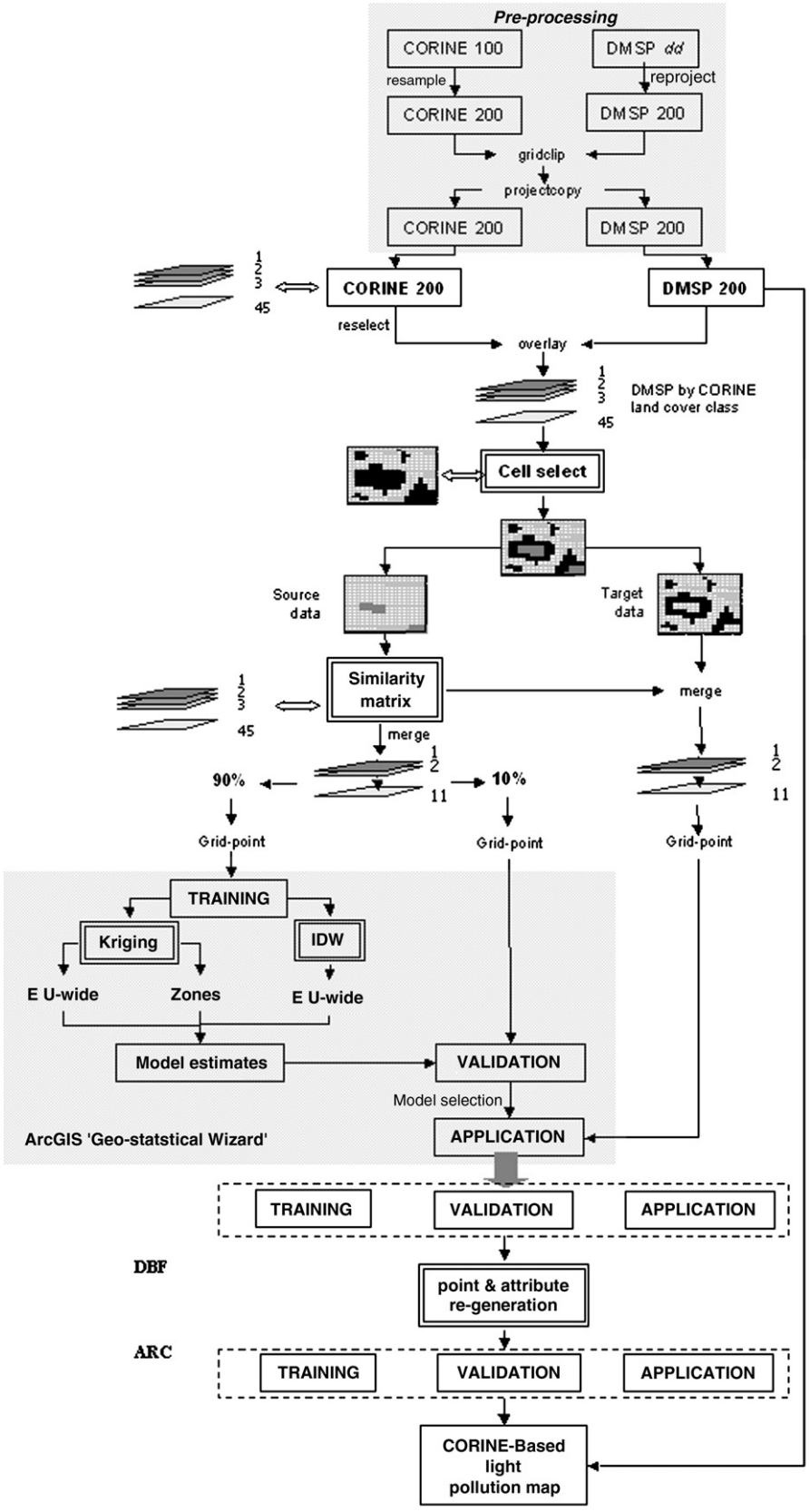

Fig. 2. Light emissions modelling strategy.

(and thus in levels of street-lighting or energy use), street-lighting technology (e.g., the use of low-reflection lighting systems), street and building configuration, and lifestyle (Elvidge et al., 1997a). For this last reason, light emissions cannot be assumed to represent a direct, linear proxy of population density; instead, local functions may need to be derived, based on an analysis of local relationships between light emissions and population density.

In order to model population density on the basis of these data, therefore, several different analytical steps are necessary. First, the light emissions data need to be sampled and analysed to reduce the effects of blooming and to obtain improved estimates of emissions at the small-area level. Second, the data need to be matched to land cover, to enable the light emissions to be redistributed more reliably to their source. Third, local relationships with land cover need to be analysed in order to derive weights applicable to specific land cover parcels.

The methodology used to achieve this is summarised in Fig. 1 and described below. All data were compiled and integrated in ArcGIS, and spatial analysis undertaken in the same environment. Statistical modelling was done in SPSS.

Table 1
Grouping of land cover classes for light emission modelling

Group name	Member classes	*N* 'training' set	*N* 'validation' set
Continuous urban fabric	1	17,188	1,890
Discontinuous urban fabric, industry, infrastructure	2, 3, 4, 5, 10	81,381	8,946
Airports, sport and leisure	6, 11	4,423	483
Construction/Dump sites	8, 9	961	111
Orchards, agro-forestry, woodland, dry heath, shrub grassland	7, 15, 21, 22, 24, 25, 26, 28, 29, 32	2,643,736	294,109
Irrigated land, pasture, salt marsh, inter-tidal	13,18, 20, 37, 39	1,828,105	203,277
Moors, annual/permanent crops, burnt areas	19, 27, 33	393,409	43,687
Rice, broad-leaf forest, dunes, marshland	14, 23, 30, 35	690,399	76,902
Non-irrigated, intensive agricultural land	12, 16, 38, 49	4,379,638	486,735
Peat bogs	36	68,934	7,685
Bare rocks and ice	31, 34	59,680	6,646

2.2. Land cover data

Land cover data were obtained from the CORINE Land Cover Map of Europe (CLC90). This was compiled on the basis of Landsat and, in some countries, SPOT imagery, using a combination of semi-automated and manual interpretation techniques (Commission of the European Communities, 1993). Source imagery derives from the late 1980s and early 1990s, but the full digital map coverage of the EU was not completed until the mid-1990s, and coverage for Sweden has never been available. For this reason, Sweden is excluded from this analysis. (A new data set, CORINE2000, based on imagery from the late 1990s and covering all 26 countries in the expanded EU is now becoming released on a country-by-country basis — see Nunes de Lima, 2005.)

CLC90 provides data on 45 land cover classes, at three levels of interpretation. Eleven of these classes relate to urban land, of which two explicitly define residential areas (continuous urban fabric and discontinuous urban fabric) — though residential properties may obviously be contained in almost all other classes. Digital data are available both in vector and raster form: for the sake of computational convenience, raster data were used here. These provide a 100-m pixel resolution, though according to the CORINE guidelines the minimum mapping unit recognised in the data is ca. 25 ha. For this study, the land cover was resampled to 200-m cell size, in order both to reduce computational demands and in recognition of the underlying limits of accuracy of the data being used.

2.3. Light emission data

Data on light emissions (from DMSP F-12) were obtained from the NOAA National Geophysical Data Center. To remove the effects of short-term variations in emissions and cloud cover, data from 10 more-or-less cloud-free nights during the winter of 1999–2000 were selected and averaged. As supplied, these data had a nominal GSD of ca. 750×600 m. This represents a notional enhancement of the original coarse resolution data produced by the currently operating sensors (2. 7 km), as a result of resampling during preprocessing. At the time of this study, data were acquired and delivered at three different gains (high, medium, low), representing different sensitivity ranges of the sensors. For the purpose of this study, these were first converted into radiances using the pre-flight calibration, and then combined using a purposely designed weighted sum procedure (developed and applied by partners at CeSIA, Italy), to give a single 32-bit measure of light emission, across all gains. Inspection of the distributions of the values in the three gains for a sample of areas suggested that:

1. The lowest values in each gain represented noise in the data;
2. At high values (digital number=63) the high gain became saturated;
3. There was overlap between the lower values in the low/medium and the high values in the high gain.

The weighted sum procedure was therefore designed (a) to stretch the distributions, (b) to trim the left (lower) tail of the distribution from the low and medium gains, and both tails from the distribution for the high gain data, and (c) to scale the adjusted values into a 16-bit number. To this end, intensity values were calculated as:

$$I = \alpha \cdot I'_L + \beta \cdot I'_M + \gamma \cdot I'_{\mathrm{H}} \tag{1}$$

where L=low gain, M=medium gain, H=high gain and:

$$I'_i = \begin{cases} I_i \text{ if } I_i \geq (\overline{I_i} - k \cdot \sigma_i) \\ 0 \text{ otherwise} \end{cases} \quad i = L,M$$

$$I'_H = \begin{cases} I_H \text{ if } I_H \in (\overline{I_H} - k \cdot \sigma_H, \overline{I_H} + k \cdot \sigma_H) \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

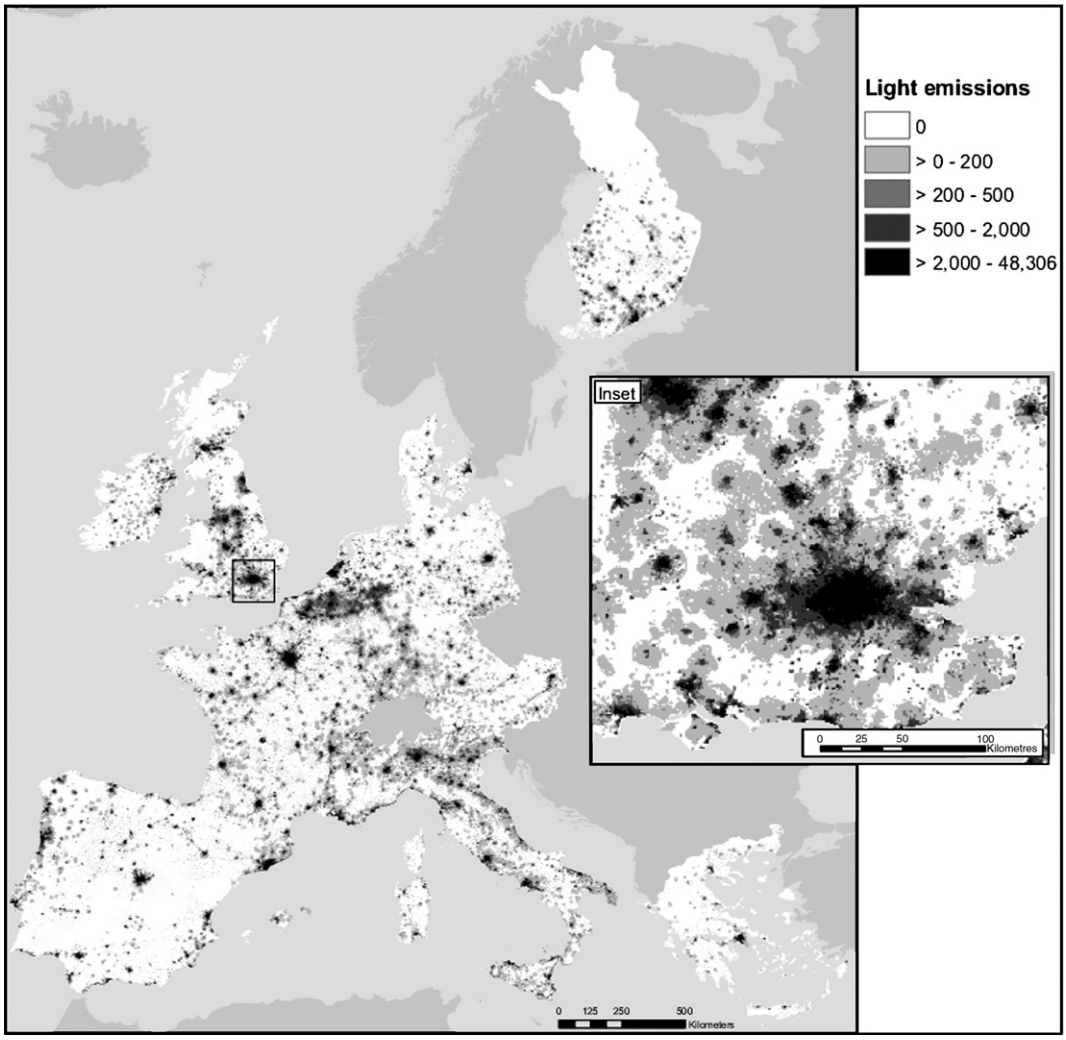

Fig. 3. Light emissions map of the EU.

where $\overline{I_i}$ and σ_i indicate the mean and standard deviation respectively of intensities measured with gain i,$k = 1.5$, $\alpha = 2^{(N-n_L)}$, $\beta = 2^{(N-L-n_M)}$, $\gamma = 1$, and $N = n_L + n_M + n_H$ is the total number of bits used to quantify the weighted intensity I.

The three intensities, I'_I, were then coded with n_i bits and organised into an N-bits word in a sequence that reflects the information content of the data: n_L, n_M, n_H.

Values of intensity are therefore based on the most informative gain available at each location, using high gain data where low and medium gain values are low (more than 1.5 SD below the overall mean for each gain) and low or medium gain data where the high gains approach saturation (>1.5 SD above the overall mean for that gain). Clouded areas were identified using the OLS thermal band and omitted from the computation, as were faulty pixels.

Light emission data were not available for northern Finland, so this was excluded from the analysis (along with Sweden). To enable registration with land cover, the data were resampled to a 200-m grid scale, and then reprojected to the Lambert Azimuthal grid.

Further processing of the data was undertaken as shown in Fig. 2. A key element of this processing was to remove blooming from the data, due to surface reflection and atmospheric scattering. Because of this, the light intensity values derived from the OLS data cannot be related directly to the underlying land cover, but represent a somewhat smoothed average across a wider area. Use of the original data is therefore likely to lead to systematic biases in the modelled population distribution, with exaggeration of the extent or urban areas and people being incorrectly assigned to areas affected by blooming from neighbouring brightly lit areas (Henderson et al., 2003; Small et al., 2005).

To address this problem, the light emissions were first reprojected into Lambert Azimuthal projection and overlaid onto land cover boundaries. The match between the two was then inspected visually along boundaries between what might be considered to be lit and unlit land cover classes (e.g., where urban areas were bounded by forest or open grassland). Based on this, it was estimated that blooming was usually confined to an area of no more than 400 m; a buffer zone of this diameter was therefore used to trim areas on either side of a land cover boundary, where blooming might have contaminated the light intensity data. To facilitate processing, the gridded data were converted to points, each representing the 200-m pixel centroids

of the land cover grid. The data were then interrogated using a moving window technique to identify all points that lay within a single land cover class, and were at least 400 m (i.e., 2 pixels) from any other land cover class. These points thus represented areas for which the attached DMSP value could be considered to provide a reliable estimate of light emissions for that land class, unaffected by blooming from adjacent classes. Data for these points were then randomly split into two subsets — one containing 90% of the selected points for each land class (to be used as a 'training' set) and the other containing the remaining 10% of selected points (to be used for validation). In order to simplify computation, the distribution of light intensities for the 45 land cover classes in CORINE were compared, and the classes combined into 11 broader groups (Table 1).

Estimates of light emissions at each of the validation locations were then made using inverse distance weighting (IDW) and universal kriging techniques in ArcGIS. For IDW, an inverse square function was applied, with a search radius (r) of 10 km, a minimum number of sample points to be used in each estimation (n_{min}) of 4, and the maximum number of contributory points (n_{max}) of 100. For kriging, a range of models was tested and compared, including linear, cubic and quadratic models, with and without global trend. Results from all models were assessed by comparing predicted light emissions with the 10% of 'pure' points retained for validation purposes.

Universal kriging slightly outperformed IDW in terms of the resulting regression statistics (R^2=0.97 and 0.96, respectively, SEE=113 and 125 respectively, for n=1,130,467 pixels). Kriging, however, occasionally produced extreme values in areas with poor data control (i.e., where there were few training points). IDW was therefore selected as the preferred method. This was therefore applied to predict light intensities at the points eliminated during buffering (i.e., those potentially affected by blooming), and a map of modelled light emissions derived. Fig. 3 shows the resulting map of Europe, aggregated to 1-km scale.

2.4. Population data

Census data are clearly the best available baseline data on population, and thus provide reference data against which any population model can be calibrated and validated. At the time of this study, the most recent high-resolution population data for Europe related to the 1991 census (even at the time of writing, small-area data for the 2001 census are not available for the whole EU). Data on 1991 population numbers were therefore obtained from the SIRE database, maintained by Eurostat. Data comprised population totals, subdivided by age and gender, for the whole of the EU at what was known as NUTS (Nationales Unites Territoriales Statistique) 5 level (now termed Local Administrative Units — LAU-2). These represent the smallest set of administrative regions for which data are consistently available; in Great Britain (GB), for example, they are represented by wards, and in France by communes.

Boundaries of NUTS 5 regions were obtained in digital form from GISCO, which serves as a warehouse for geographical data in the EU. Boundaries dated from 2001. Some problems were consequently encountered in matching population data to these boundaries, largely because of changes in administrative geography of the member states since the 1991 censuses. For this reason, only NUTS level 3 data (equivalent to Kreis) were available for eastern Germany, and 2003 population data at NUTS 5 level had to be used in western Germany. For the UK, the 1991 boundaries and population counts obtained from the national data sources (see below) were substituted for those from GISCO and SIRE in order to facilitate *post hoc* validation against national data.

Details of the resulting database are summarised in Table 2.

2.5. Population modelling

Modelling was done using regression analysis techniques in SPSS, with the NUTS 5 level population count as the dependent variable, and the area of lit land, area of unlit land and total light emissions for each land class type in each NUTS 5 region as predictor variables. Weights derived from these models were then used to redistribute the NUTS 5 level population totals to each pixel pyconphylactically. To enable this, the point coverage of modelled light emissions was first intersected with the NUTS 5 boundaries. For each of the 45 land cover classes, three variables were then computed within each NUTS 5 region: NU, the number of points for which the light emission was zero (equivalent to area of unlit land); NL, the number of lit points (equivalent to the area of lit land); and LE, the total light emissions.

Four different modelling strategies were applied, with different aggregations (groups) of the original land cover classes, as shown in Table 3. These were designed to represent increasingly simplified classifications, from strategy 1 to strategy 4, in order to determine how robust the models were to the spectral resolution of the land cover data. In each case regression models were constructed in a supervised, stepwise process, entering groups of variables (for lit area, unlit area and

Table 2
Summary statistics for census data

Country	N	Area		Population	
		Mean	SD	Mean	SD
Austria	2358	35.6	37.5	3312.8	32738.1
Belgium/Luxembourg	708	46.9	36.8	14636.5	26367.4
Denmark	277	156.2	101.3	18579.3	36070.5
Finland	434	721.5	832.5	11190.3	30265.2
France	36570	15.0	15.3	1548.5	13938
Great Britain[1]	10528	21.8	61.9	5204.1	3894.9
Germany (East)[2]	114	953.3	736.3	153948.4	317426.4
Germany (West)[3]	8789	28.3	34.1	7458.4	35468.9
Greece	1034	127.9	107.7	9938.0	28216.5
Ireland[4]	3990	21.2	18.8	1262.9	1550.5
Italy	8109	37.2	50.8	6998.1	42428.6
Netherlands	505	82.1	91.2	29447.9	51724.1
Portugal[5]	4007	22.3	36.1	2335.5	4833.3
Spain[6]	8094	61.7	93.0	4603.9	43768.9
EU[7]	85403	31.0	91.9	3965.9	25480.9

Notes: (1) based on 1991 census and ward data; (2) based on NUTS 3 regions; (3) based on 2003 population data; (4) includes N. Ireland; (5) excludes Madeira; (6) excludes Canary Islands; (7) excludes E. Germany.

Table 3
Land cover categories, and order of entry, for each modelling strategy

CORINE land cover class(es)	Strategy 1		Strategy 2		Strategy 3		Strategy 4	
	Order	Category	Order	Category	Order	Category	Order	Category
1	2	Continuous urban	2	Continuous urban	2	Continuous urban	1	Residential
2	1	Discontinuous urban	1	Discontinuous urban	1	Discontinuous urban		
49	3	Unclassified surfaces	3	Unclassified surfaces	3	Unclassified surfaces	2	Unclassified surfaces
3	5	Industrial/commercial	5	Industrial/Commercial	5	Non-residential urban	3	Non-residential urban
10	8	Institutions	7	Open urban				
11	9	Urban green space						
4–6	10	Transport	8	Other urban				
7–9	13	Waste, extraction and construction sites						
12–17, 19	4	Arable and permanent crops	4	Arable and permanent crops	4	Cultivated	4	Rural
18	6	Pasture	6	Pasture				
20	11	Complex cultivation patterns	10	Other cultivated				
21	12	Land principally occupied by agriculture						
23	7	Broad-leaved woodland	9	Woodland	6	Uncultivated		
22, 24, 25	14	Other woodland						
26–29	15	Unimproved grass and scrub	11	Unimproved grass and scrub				
30–38	16	Other uncultivated	12	Other uncultivated				

light intensity) in the order shown. This order is important, because colinearity between several of the land cover classes means that they compete to explain the variations in population density, creating instability in the models. The entry sequence shown in Table 4 was thus defined, *a priori*, to reflect the likely population density of the land cover groups — an assumption that can be assessed by comparing the slope coefficients derived from the regression analysis.

Variables were retained only if they had positive coefficients and were significant at the 0.05 confidence level. In order to maintain the pre-defined hierarchy of importance of land cover groups, new variables were also allowed to enter the model only if they did not cause the removal of entire land cover groups entered earlier in the analysis, though individual variables (e.g., NU, NL or LE) could be removed if their significance fell below $p=0.1$. For the final model, also, an additional criterion was specified, that the constant should also be positive. By applying these criteria, therefore, the models selected were not necessarily those that gave the best prediction of population at the NUTS 5 level, but instead were constrained to avoid the possibility of negative population predictions for any location, to avoid counter-intuitive relationships (e.g., negative associations with light emissions) and to provide estimates of 'background' population densities in land cover classes not incorporated into the model.

The resulting models thus took the general form:

$$\mathrm{PW}_k = \mathrm{PB}_k + \sum_{j=1}^{n} \left[(l.\mathrm{NL}) + (u.\mathrm{NU}) + (e.\mathrm{LE})\right]_j \quad (3)$$

where: PW_k=population in NUTS area k; NL_j=number of lit pixels of land cover group j in NUTS area k; NU_j=number of unlit pixels of land cover group j in NUTS area k; LE_j=total light emission from land cover group j in NUTS area k; l, u and e=variable-specific weights (regression coefficients); PB_k='background' population (as defined by the constant).

Table 4
Regression statistics for light emissions models

	Strategy 1			Strategy 2			Strategy 3			Strategy 4			Population threshold
	R^2	SEE	SEE/Mean	R^2	SEE	SEE/Mean	R^2	SEE	SEE/Mean	R^2	SEE	SEE/Mean	
AT	0.90	2581	0.78	0.90	2702	0.82	0.92	2362	0.71	0.89	2740	0.83	1,500,000
BE-LU	0.87	7255	0.50	0.86	7451	0.51	0.81	8708	0.59	0.83	8267	0.56	400,000
DE-east	0.84	25414	0.17	0.83	26200	0.17	0.83	26002	0.17	0.82	26712	0.17	3,000,000
DE - west*	0.97	4246	0.57	0.97	4670	0.63	0.97	4611	0.62	0.96	5160	0.69	900,000
DK	0.95	4071	0.22	0.95	4097	0.22	0.95	4367	0.24	0.91	5606	0.30	260,000
ES-PT	0.81	7886	2.05	0.80	8117	2.11	0.80	8140	2.11	0.74	9252	2.40	1,600,000
FI	0.97	5459	0.49	0.97	5555	0.50	0.97	5367	0.48	0.94	7146	0.64	NA
FR	0.83	2794	1.80	0.82	2848	1.84	0.82	2862	1.85	0.80	3060	1.98	400,000
GB	0.73	2033	0.39	0.73	2032	0.39	0.73	2037	0.39	0.72	2050	0.39	NA
GR	0.82	6644	0.67	0.81	6835	0.69	0.81	6889	0.69	0.70	8660	0.87	300,000
IE-NI	0.73	802	0.63	0.73	801	0.63	0.73	813	0.64	0.72	815	0.65	NA
IT	0.86	7386	1.06	0.85	7449	1.06	0.85	7611	1.09	0.81	8387	1.20	900,000
NL	0.92	7946	0.27	0.92	7576	0.26	0.92	7632	0.26	0.92	7968	0.27	200,000
EU	0.82	4518	1.14	0.81	4551	1.15	0.81	4634	1.17	0.77	5063	1.28	300,000

Notes: * Modelled using 2003 population data for western DE.

Table 5
Light emission models: strategy 2

Land cover class	Index	EU	AT	BE-LU	West DE	East DE	DK	ES-PT	FI	FR	GB	GR	IE-NI	IT	NL
Constant		87.3	140.2	2368.3	427.5	42977.5	3413.9	125.0	2150.2	82.4	1597.2	1586.7	414.0	30.0	1087.6
1: Continuous urban	Nl	187.5		470.9	988.4	770.2	545.7	251.7	366.2	394.0	167.3	664.9	116.3		192.0
	Nu	140.2	1288.1		1205.9						161.2			200.0	178.6
	Le	0.023						0.049		0.013	0.003		0.002	0.042	0.007
2: Discontinuous urban	Nl	100.7	176.6		106.5	39.1	49.9	115.6			115.0	66.7	98.7	129.9	52.3
	Nu	84.8	71.1			37.8	25.0	75.5		52.8	71.1	94.8	100.7	121.7	112.2
	Le	0.007	0.003	0.026	0.018	0.036	0.013		0.005	0.027	0.008	0.012	0.006	0.012	
3: Industrial/commercial	Nl	145.5	229.9	212.1	86.9	110.8	97.5	303.9	65.3	162.3				63.3	129.9
	Nu	95.0	234.8	327.4			213.8	197.4		99.3		49.9		42.2	
	Le				0.064								0.005		
4–9: Other urban	Nl	9.1		26.4				94.2	32.9	19.6			15.1		
	Nu	11.0			61.1								13.3	92.6	
	Le	0.006					0.010		0.011					0.054	0.046
10, 11: Open urban	Nl	119.3		45.3				582.1				209.6			
	Nu	24.0		149.9	59.7				2941.5	115.4				129.2	
	Le		0.170		0.050		0.024			0.020	0.006		0.001	0.074	0.049
12–17, 19: Arable and permanent crops	Nl	0.6						3.1	2.3			5.1			
	Nu	0.2											1.1		
	Le	0.005			0.010			0.008		0.005		0.012		0.013	
18: Pasture	Nl	1.3	5.7							0.9					
	Nu	0.5	3.3		2.8		4.9						0.2		
	Le			0.035				0.049					0.001		
20, 21: Other cultivated	Nl		5.9									12.1	1.3		
	Nu				8.0					1.3		1.0	0.9		
	Le	0.015			0.093				0.094	0.003		0.014		0.018	
22–25: Woodland	Nl						10.0	1.7		1.2				5.2	
	Nu	0.0			1.8				0.1						
	Le		0.010		0.059					0.006				0.007	
26–29: Unimproved grass and scrub	Nl							6.5		1.2	1.1				
	Nu														
	Le	0.009							0.027		0.012		0.047		
30–38: Other uncultivated	Nl														
	Nu	0.4						0.8							
	Le							0.021	0.564						
49: Unclassified surfaces	Nl	18.2								54.4		28.8	53.1		
	Nu	1.0									12.3			52.9	
	Le							0.484							
Adj.R^2		0.813	0.895	0.863	0.967	0.828	0.953	0.796	0.966	0.822	0.728	0.812	0.733	0.852	0.923
Standard error		4551	2702	7451	4669	26200	4096	8116	5555	2848	2031	6834	801	7448	7576

The models obtained were then applied to calculate population numbers (PM) for each 200-m pixel within the study area, by redistributing the total population (PC, from the census) in each NUTS area to the constituent pixels. Pixels were classified (0 or 1) according to three types for this purpose: *lit* (i.e., pixels with light emissions in land classes included in the model), *unlit* (i.e., pixels with no light emissions, but in land classes included in the model), and *background* (i.e., pixels in land classes not otherwise included in the model). Populations were then computed as follows:

$$\mathrm{PM}_{ijk} = \frac{\mathrm{PC}_k.[(l.\mathrm{lit}) + (e.\mathrm{LE}) + (u.\mathrm{unlit}) + (\mathrm{background}^*\mathrm{Pb}_k/\mathrm{Nb}_k)]}{\sum_{i=1}^{n} [(l.\mathrm{lit}) + (e.\mathrm{LE}) + (u.\mathrm{unlit}) + \mathrm{background}^*\mathrm{Pb}_k/\mathrm{Nb}_k)]_i} \quad (4)$$

where Nb_k=the number of background pixels in NUTS area k (i.e., all pixels not otherwise included in the model).

Thus the modelled population was rescaled pycnophylactically to match the actual (census-derived) population in each NUTS area.

Models were developed at both national level, and for the EU as a whole (EU models excluded East Germany because of the different NUTS level of the population data). In each case, distributions of the population data were first explored to identify outliers that might bias the regression models. Thresholds were then selected to exclude these NUTS areas. The number of excluded areas ranged from 0 (three countries) to 4 (Italy). Predictions for these areas were then made from the model. As a basis for comparison, models were also developed for each strategy and geographic area using land cover area alone (i.e., without light emissions data). Results of the different models and modelling strategies were assessed and compared by reference to the regression statistics (R^2, SEE and SEE/mean) and then further validated by comparison with independent, small-area population data in Great Britain.

3. Results

Results of modelling using the light emissions data are summarised in Table 4, and details of the models for Strategy 2 are given in Table 5. As Table 4 shows, all the modelling strategies performed relatively well at both national and EU level, with R^2 generally between 0.8 and 0.9, and SEE in the range of 2–5000 people. Differences between strategies were

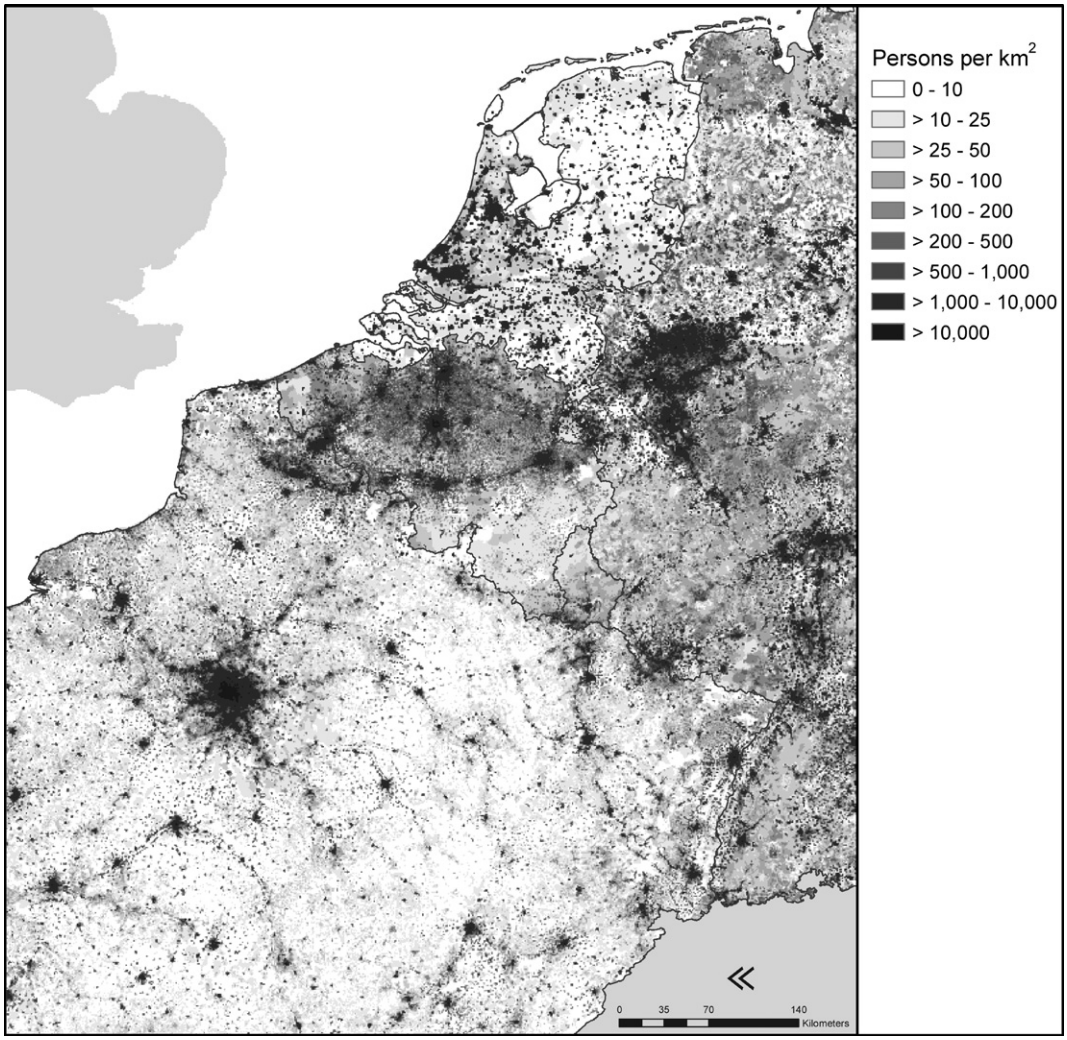

Fig. 4. Population density: an example of disparities in national models — France– Belgium–Netherlands border region (strategy 2).

slight, though overall performances declined slightly as the models were simplified (from strategy 1 to strategy 4).

Some differences in model performance (as indicated by the regression statistics) are apparent between countries. The lowest R^2 values (~0.7) are seen in Great Britain and Republic of Ireland/Northern Ireland. In part, this possibly reflects the way in which the administrative areas in these countries are designed to limit variations in population totals, such that the NUTS populations on which the models are built vary relatively little (see Table 2). Reflecting this, the SEE in both sets of countries is small. The large SEE values seen in East Germany, in contrast, reflect the necessity to use population data from NUTS 3 regions, rather than NUTS 5.

The structure of national models is broadly similar across all countries: data for strategy 2 (Table 5) are typical in this respect. In most cases, the numbers of lit pixels and light emissions for the two main residential classes (continuous urban and discontinuous urban) are included and dominate the model (typically accounting for between 70 and 90% of the total R^2). Unlit areas of these land classes also appear in the model in a number of cases. The numbers of lit and unlit pixels of industrial and commercial land likewise appear in many of the models, as do lit areas (and light emissions) for areas classified as open urban land. The role of other (including rural) land classes is more variable, though these rarely account for more than 10–20% of the overall variation in modelled population.

Coefficients for these various parameters are generally logical, and follow expected variations in population density. Weights (i.e., implied population densities) for continuous urban areas, for example, are typically higher than those for discontinuous or other urban land cover groups, and much greater than those for rural areas. Similarly, coefficients for lit areas are usually greater than for unlit areas, especially for urban land cover groups. The intercept value (constant), which is taken to represent the 'background' population, varies substantially: as might be expected, it tends to be low in countries where several rural land classes enter the model, and higher in those where the number of rural land classes is small.

A single EU-wide model was also developed and is reported in Tables 4 and 5. In terms of R^2, this performs less well than the median of the national models (0.77–0.82 across the four strategies compared with a median of 0.82–0.87), but the SEE is

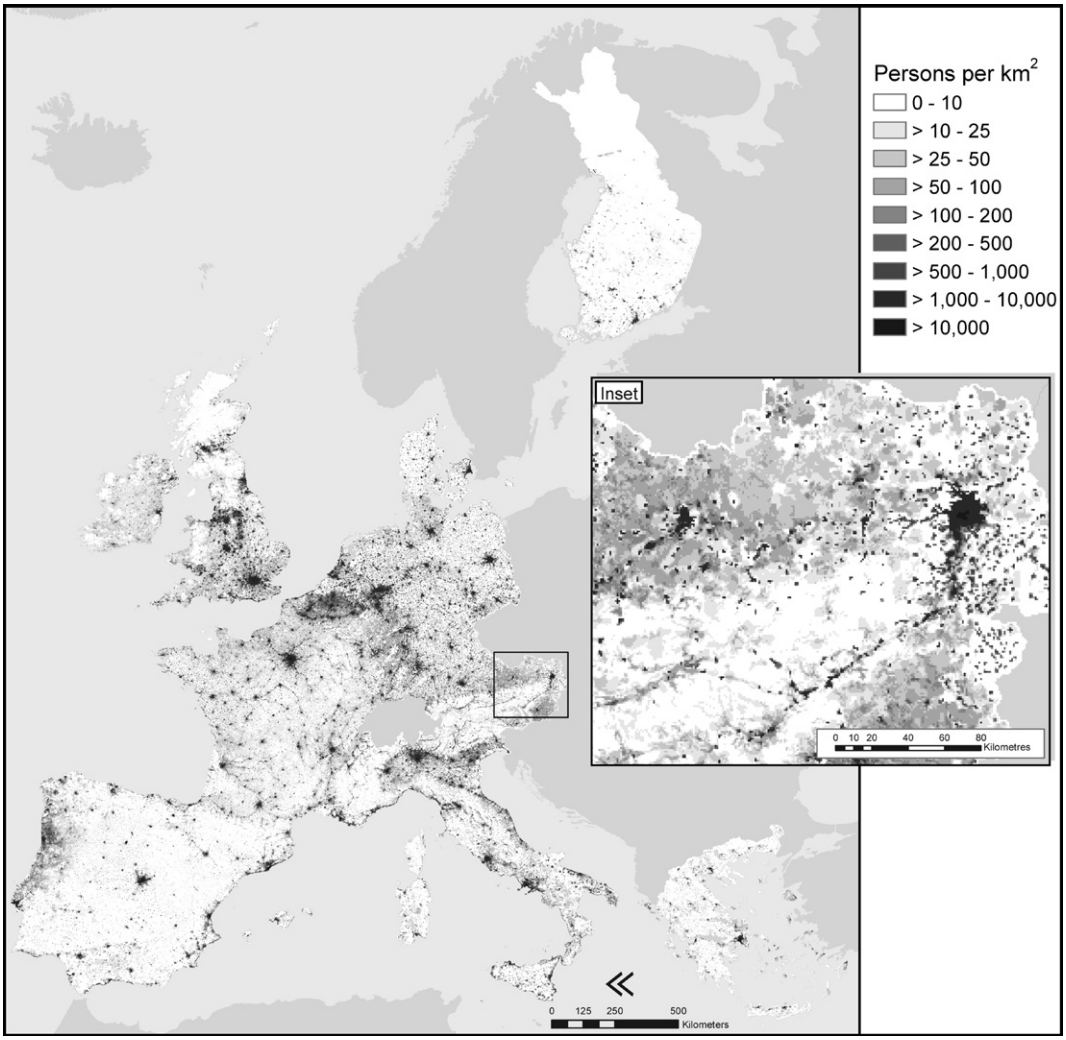

Fig. 5. Population density: EU model (strategy 2).

slightly smaller than the national median — 4518–5063 compared with 4852–6376, (excluding East Germany).

Both national and EU models were also applied to compute population numbers for each 200-m pixel, using Eq. (3) above, and the results then aggregated to 1 km for mapping. National maps were in every case realistic, which is to be expected given that the modelled populations are conditioned to tally with the NUTS 5 totals, but when mapped across the EU the results in some cases showed marked discontinuities at national borders — e.g., between France, Belgium and The Netherlands (Fig. 4 shows the example for strategy 2). These discontinuities were not visible in maps based on the EU model. For EU-wide population modelling, therefore, the EU model is preferred. Fig. 5 shows the example from strategy 2.

As noted, models were also developed using only the area of each land cover class (i.e., without light emissions). Models for Austria, Finland and The Netherlands all perform well, with R^2 in excess of 0.9 and standard errors small compared to the mean population size of the NUTS regions (Table 6). For Spain and Portugal, Greece and France, in contrast, the models are relatively poor, with R^2 in the range of 0.4–0.55 and standard errors substantially in excess of the average population size. For most countries, model performance (as indicated by the regression statistics) tends to decline as the models are simplified — in some cases (e.g., Spain/Portugal, Finland, France, Italy and East Germany) quite markedly. Instability is also evident in some national models: in both Denmark and Belgium/Luxembourg R^2 recovers for strategy 4. The EU model is again intermediate, with R^2 between 0.72 and 0.77 and a standard error of 5038–5560 (compared to an average NUTS population of 3966).

Comparisons between the results in Tables 4 and 6 are informative. They suggest that models based only on land cover perform as well as those with light emissions in several northern and western countries, and in mountain areas such as Austria, but markedly worse in the south. The reasons for this are not clear. It may reflect differences in the way the CORINE land cover classification has been applied in different member states. It might also reflect inherent differences in the relationships between land cover and light emissions in different parts of Europe — for example, because of differences in parcel size and heterogeneity of urban land classes, differences in building characteristics or surface materials and hence reflection properties, or differences in lighting technology and practice. Whatever the cause, the results suggest that there are advantages in using light emissions as part of the modelling process for EU-wide analysis. This interpretation is supported by the differing sensitivities of the land cover and light-based models to the different modelling strategies. Inclusion of light emissions in the models (Table 4) seems to allow for local differences in population density, even when the land cover categories themselves are aggregated. When light data are not used (Table 6), more detailed land cover classifications are needed to provide adequate discrimination of variations in population density.

3.1. Post hoc validation

3.1.1. Data acquisition and preprocessing

Opportunities for independent validation of the population models are limited by the lack of suitable reference data for most countries. Detailed cadastral and census data do exist for several countries, notably The Netherlands and Scandinavia, but these are not readily available for research purposes. Post hoc validation was possible in Great Britain, however, using the detailed population counts provided by the Office for National Statistics. For 1991, data are available for enumeration districts (ED, $n=147{,}596$); for 2001 they are available both for output areas (OA, $n=218{,}038$) and postcodes ($n=1.38$ million).

For the purpose of validation, modelled population data had to be transformed to the same geographic units as the reference counts. Modelled populations (by 200-m pixel) from both the GB national and EU models were therefore totalled to match the 1991 EDs and 2001 OAs: to avoid the large computational demands involved in polygon overlay, this was done by

Table 6
Regression statistics for land cover area models

	Strategy 1			Strategy 2			Strategy 3			Strategy 4			Population threshold
	R^2	SEE	SEE/Mean	R^2	SEE	SEE/Mean	R^2	SEE	SEE/Mean	R^2	SEE	SEE/Mean	
AT	0.95	1954	0.59	0.92	2328	0.70	0.91	2502	0.76	0.87	2970	0.90	1,500,000
BE-LU	0.75	10160	0.69	0.73	10387	0.71	0.39	15703	1.07	0.61	12633	0.86	400,000
DE-east	0.77	30410	0.20	0.74	32227	0.21	0.73	32913	0.21	0.45	46974	0.31	3,000,000
DE - west*	0.92	7461	1.00	0.90	8186	1.10	0.86	9790	1.31	0.82	10872	1.46	900,000
DK	0.81	8210	0.44	0.81	8303	0.45	0.64	11274	0.61	0.84	7607	0.41	260,000
ES-PT	0.40	13954	3.62	0.25	15529	4.03	0.23	15755	4.09	0.05	17558	4.56	1,600,000
FI	0.96	6258	0.56	0.95	6555	0.59	0.97	5084	0.45	0.60	19208	1.72	NA
FR	0.56	4486	2.90	0.55	4556	2.94	0.43	5122	3.31	0.01	6729	4.35	400,000
GB	0.71	2095	0.40	0.71	2107	0.40	0.71	2107	0.40	0.69	2156	0.41	NA
GR	0.42	11977	1.21	0.42	12020	1.21	0.38	12369	1.24	0.37	12508	1.26	300,000
IE-NI	0.72	822	0.65	0.72	824	0.65	0.72	827	0.65	0.71	832	0.66	NA
IT	0.71	10421	1.49	0.69	10837	1.55	0.56	12828	1.83	0.45	14367	2.05	900,000
NL	0.92	7777	0.26	0.91	8352	0.28	0.90	8478	0.29	0.89	9012	0.31	200,000
EU	0.77	5038	1.27	0.77	5057	1.28	0.76	5165	1.30	0.72	5560	1.40	300,000

Notes: * Modelled using 2003 population data for western DE.

attaching the modelled populations to the centroid of each pixel, and carrying out a point-in-polygon overlay. In addition, comparisons were made for a regular 1 km grid. For 1991, 1 km totals for the reference populations (PR) were derived from the enumeration district (ED) totals by postcode weighting, as follows:

$$PR_k = \sum_{i=1}^{n} \frac{PE_i \cdot p_{ik}}{p_i} \quad (5)$$

where: PE_i=population in ED i; p_{ik}=number of postcodes in ED i in grid cell k; p_i=number of postcodes in ED i.

Postcode locations were obtained from the 1991 All Fields Postcode Directory (AFPD) maintained by the Post Office, but, because of known errors, were subjected to extensive checking and revision by comparison with later postcode files before use. Intersection between the corrected postcode locations and EDs was done in ArcGIS. For 2001, 1 km population counts were obtained by summing the postcode headcount data, based on an intersection of postcode locations from Codepoint 2001.

Four sets of population data were thus compiled for validation purposes, as follows:

- ED1991 — enumeration district totals, derived from the 1991 census,
- KM1991 — 1 km totals for 1991, obtained by postcode weighting of ED1991 data to a 1 km grid,
- OA2001 — output area totals, derived from the 2001 census,
- KM2001 — 1 km totals for 2001, obtained by summing the 2001 headcount data to a 1 km grid.

3.1.2. Model comparison and evaluation

Results for the different validation studies are shown in Table 7. Two measures of performance are reported: R and the root–mean–square error (RMSE). In general, the GB national model performs marginally better than the EU model. At the 1-km scale, an extremely strong correlation is evident for all strategies, and for both census years, with R=0.90–0.96 and RMSE in the order of 300 (compared with a mean population of 218). For ED1991 the correlations were much weaker (0.33–0.40), and RMSE values relatively high (~350). At the yet finer output area level, model performance declines further, with R=0.23–0.27 and RMSE ca. 275–300.

In interpreting these results, a number of factors must be considered. Differences in the dates of the various datasets used clearly represent an important source of uncertainty, though changes in population in Great Britain between 1991 and 2001 (like the rest of the EU) have generally been small, so these effects are likely to be limited. The population data used for validation at the 1-km scale are also subject to potential error, due to the way in which the counts have been spatially transformed using postcode locations or headcounts.

Validation at the ED and OA faces further difficulties because of the small size of these units: for 1991 EDs, the median area (and inter-quartile range) was 0.10 km^2 (0.04–0.35), for 2001 OAs it was 0.06 km^2 (0.03–0.15). At their smallest, therefore, these units are finer than the 200-m (0.04 km^2) cell size used in the population modelling. Because they are also designed to have more-or-less equal populations, both EDs and OAs are especially tiny in densely populated urban areas. A small percentage (0.8%) of EDs are also false, in that they contain no population, whilst a few EDs represent large institutions (e.g., prisons) with very limited spatial extent. These therefore constitute very challenging targets for modelling, and it is likely that, in many instances, pixel centroids became attached to the incorrect administrative area. At the same time, modelling in rural areas faces the problem of predicting the location of small, often scattered settlements across large land cover parcels and large EDs or OAs. The effects of these factors are indicated by analysis of the residuals from the population modelling at output area level (Fig. 6). For output areas less than 0.01 km^2 in area, the absolute residual is relatively large, even though the average population of these areas is small (233). For output areas between 0.01 and 0. 2 km^2, the residual is smaller, though more variable; for the larger, mainly rural OAs (>0. 2 km^2) it increases again, with greater variability.

Overall, therefore, the results of *post hoc* validation suggest that the models provide a very good estimate of population number in Great Britain at 1 km level. At finer resolutions (e.g., at ED or OA), and especially at the scale of the smallest administrative area (<0.01 km^2), the estimates are less reliable.

Table 7
Results of validation analyses

Validation data set		Target population		Strategy 1		Strategy 2		Strategy 3		Strategy 4	
File	N	Mean	SD	R	RMSE	R	RMSE	R	RMSE	R	RMSE
GB models											
ED1991	147596	370	191	0.391	346	0.391	345	0.394	343	0.397	342
KM1991	238790	228	835	0.958	239	0.958	238	0.959	239	0.959	239
OA2001	218034	239	134	0.264	276	0.264	277	0.265	275	0.266	274
KM2001	263101	218	838	0.951	265	0.951	264	0.951	263	0.951	263
EU models											
ED1991	147596	370	191	0.334	412	0.334	411	0.343	401	0.353	397
KM1991	238790	228	835	0.938	290	0.938	281	0.942	276	0.944	240
OA2001	218034	239	134	0.234	320	0.233	321	0.235	316	0.240	312
KM2001	263101	218	838	0.931	307	0.931	306	0.935	298	0.937	293

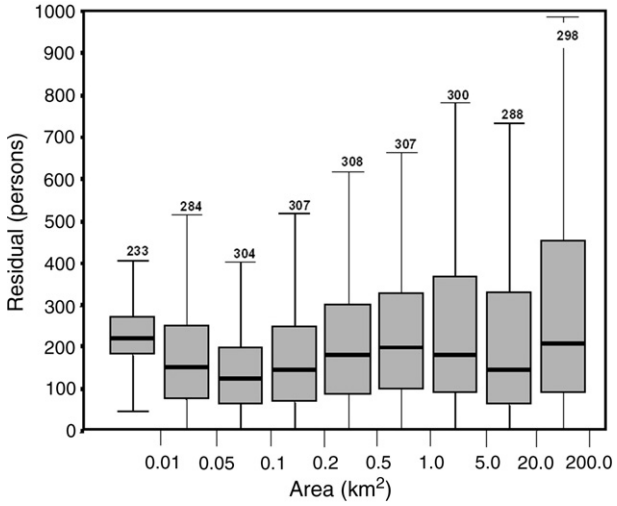

Fig. 6. Boxplots of absolute residuals from the GB model (strategy 2) by size of output area.

Because model performance is somewhat poorer in Great Britain than other countries, this probably represents a worst-case condition. In other countries, where the regression models are seen to be intrinsically more robust at the NUTS 5 level, reliability of the modelled estimates might be expected to be better.

4. Discussion

Despite efforts to improve and standardise population census procedures over recent decades, difficulties clearly remain in obtaining reliable small-area population estimates in many parts of the world. Even when the population counts are reported, access to the geographical information (e.g., census tract boundaries) needed to analyse and map the data is often difficult. A coherent set of detailed population counts and matched boundary files for the 2001 census across the EU, for example, is not available, so any analysis would involve collating the population counts and boundary data from each country and building a European coverage (no simple task because of the inevitable lack of consistency between these files both within and between countries). Where international comparisons are necessary, as across different members of the EU, differences in the size and character of the census tracts also pose problems. In addition, for many applications, there is a need to transform population data between different zone systems, in order, for example, to obtain denominators for health outcome data, to estimate population exposures to environmental pollutants, or to derive indices of population pressures on the environment. There is thus a continuing need for methods for small-area population modelling.

With the development of remote sensing and GIS techniques, opportunities for such modelling have been greatly enhanced. Simple, traditional approaches of area-weighting can therefore be improved, using land cover data. In order to make maximum use of land cover data, however, methods are needed that enable the derivation of parcel (or even sub-parcel) level indicators of variations in population density. The approach presented here, using light emission data, offers one such methodology. Based on the results obtained, this would seem to provide a basis for reliable population modelling to a resolution of at least 1 km, and in many circumstances possibly finer scales.

That said, the limitations of the methods used here need to be recognised. Amongst these, statistical issues are of particular note. In common with many geographical analyses (Anselin, 1989; Haining, 1990, 1991), the data used here do not conform fully to the demands of regression analysis, in that they often depart from conditional normality. Consideration was given to normalising the data (e.g., by log-transformation), which would also have resolved the problem of negative population predictions. The large proportion of zero values for several of the independent variables used for population modelling nevertheless meant that this could not fully resolve the distributional problems of the data, and was also found in many cases to lead to illogical weights in the regression models. Use of index values to transform the independent variables was also considered to be inappropriate, since it would have resulted in substantial loss of information. Modelling was therefore carried out with un-normalised variables, whilst acknowledging that the goodness of fit measures (e.g., R^2 and SEE) thereby produced are liable to be somewhat unreliable. No account is also taken in the models used here of spatial auto-correlation: use of Bayesian techniques to take account of spatial structure in the data was not computationally feasible with the large data sets involved in this analysis. Again, this is likely to mean that model performance is somewhat over-estimated. It needs to be stressed, however, that the purpose of regression analysis was not to develop models that were then used directly to predict population numbers, nor to test hypotheses about spatial relationships, but rather to derive weights for each land cover class in each NUTS area that could be used to redistribute the population totals. This pycnophylactic rescaling of the modelled estimates to match the small-area population counts means that any errors are intrinsic to each NUTS area. Independent validation of the modelled results in the UK suggests that the results are broadly reliable, at least to a spatial resolution of ca. 1 km.

Uncertainties in the relationships between light emissions and population distribution or density also need to be recognised. Previous studies exploring and exploiting these relationships at national level (Elvidge et al., 1997a,b; Sutton et al., 2001) have reported strong associations between lit area and population number, but they have also shown that light emissions depend on affluence (e.g., as expressed by GDP) and economic structure (e.g., degree of industrialisation). At the small-area level analysed here many other factors are also likely to intercede, including differences in urban configuration, transport infrastructure, energy and lighting policies (and their level of implementation) and lighting technology. In many western cities, commercial advertising, sports facilities and security lighting represent additional, though often local, sources of light emissions. For all these reasons, light emissions do not translate directly or consistently into population distribution. In this study, linkage of the light emissions data to land cover information helped to reduce some of these effects (e.g., by reducing the weight given to

commercial and industrial areas), and this certainly improved the predictions from the models. Interpolation of the light emissions data also acts to smooth out local lighting hotspots (e.g., associated with sports stadia) and may thus improve local estimates. Nevertheless, differences in the form of the national models, and the fit of the EU model, between countries suggest that residual uncertainties remain. Further analysis of some of the regional variations in model performance may give indications of possible confounding factors, and provide a basis for improving the models by incorporating additional variables (e.g., GDP, building heights). The limited availability and quality of such data need, however, to be recognised.

Other issues are circumstantial, and could be avoided or reduced in the future as improved data become available. Population modelling was carried out using the 1991 population data (for most countries), land cover data from about the same date, and light emissions data derived from 1999–2000. The mismatch in the dates of the data sets used inevitably creates temporal ambiguity in the modelled populations and generates errors in the population models. Improved estimates (and more interpretable results) could be achieved when data for a common time period become available. The spatial resolution of both the CORINE land cover data (notionally 100 m but resampled here to 200 m) and the light emissions data (originally 2.7 km, but resampled and remodelled to provide notional 200-m resolution) is also sub-optimal. The new land cover data for the EU, now being released, will improve on these resolutions, and new generation satellites (e.g., Envisat) will provide further improvements in the future. The analysis undertaken here was also restricted by the resolution of the light emissions data available to the study, following problems with the on-board storage devices in the mid-1990s. In recent years, fine resolution data supply has been resumed. Use of these data would further enhance the capability for population modelling. In the future, therefore, the opportunity to improve on the results obtained here will undoubtedly arise. In the meantime, the methods used, and population estimates obtained, are offered for wider use. Both the input data and results may be obtained from the authors.

Acknowledgements

The work reported here was undertaken as part of several different studies, including the EU funded MANTLE and APMoSPHERE projects. The financial support offered by the EU through these projects is gratefully acknowledged. Thanks are also due to the other members of these projects who contributed to the work, notably Graham Deane (Huntings Development plc), M. Benvenuti and C. Dambra (CeSIA, Accademia dei Georgofili), C. Conese (CNR–IATA), Mike Petrakis (National Observatory of Athens) and Gavin Shaddick (University of Bath).

References

Anselin, L. (1989). *What is special about spatial data? Alternative perspectives on spatial data analysis. Technical Report 89-4*. Santa Barbara, California: National Center for Geographic Information and Analysis. University of California.

Boyle, P., & Dorling, D. (2004). Guest editorial, the 2001 UK census, remarkable resource or bygone legacy of the 'pencil and paper era'? *Area*, *36*(2), 101−110.

Chen, K. (2002). An approach to linking remotely sensed data and areal census data. *International Journal of Remote Sensing*, *23*, 37−48.

CIESIN (Center for International Earth Science Information Network). (2000). *Gridded population of the world. Vers. 2*. Palisades, New York: Columbia University.

Cinzano, P. (2000). The growth of light pollution in north-eastern Italy from 1960–1995. In P. Cinzano (Ed.), *Measuring and modelling light pollution. Memoirs of the Society of Astronomy of Italy*, *Vol. 71*. (pp. 159−166).

Commission of the European Communities. (1993). CORINE Land Cover technical guide. European Union. (Luxembourg, Directorate-General for the Environment, Nuclear Safety and Civil Protection).

Cook, L. (2004). The quality and qualities of population statistics, and the place of the census. *Area*, *36*, 111−123.

Elvidge, C. D., Baugh, K. E., Kihn, E. A., Kroehl, H. W., Davis, E. R., & Davis, C. W. (1997). Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. *International Journal of Remote Sensing*, *18*, 1373−1379.

Elvidge, C. D., Baugh, K. E., Kihn, E. A., Kroehl, H. W., & Davis, E. R. (1997). Mapping city lights with nighttime data from the DMSP Operational Linescan System. *Photogrammetric Engineering and Remote Sensing*, *63*, 727−734.

Flowerdew, R., & Green, M. (1994). Areal interpolation and types of data. In S. Fotheringham, & P. Rogerson (Eds.), *Spatial analysis and GIS* (pp. 121−145). London: Taylor and Francis.

Goodchild, M. F., & Lam, N. S. -N. (1980). Spatial interpolation methods, a review. *American Cartographer*, *10*, 129−149.

Haining, R. (1990). *Spatial data analysis in the social and environmental sciences.* Cambridge: Cambridge University Press.

Haining, R. (1991). Estimation with heteroscedastic and correlated errors, a spatial analysis of intra-urban mortality data. *Papers in Regional Science*, *70*, 223−241.

Harris, R. J., & Longley, P. A. (2000). New data and approaches for urban analysis, modelling residential densities. *Transactions in GIS*, *4*(3), 217−234.

Henderson, M., Yeh, E. T., Gong, P., Elvidge, C., & Baugh, K. (2003). Validation of urban boundaries derived from global night-time satellite imagery. *International Journal of Remote Sensing*, *24*, 595−609.

Imhoff, M. L., Lawrence, W. T., Stutzer, D. C., & Elvidge, C. D. (1997). A technique for using composite DMSP/OLS 'city lights' satellite data to map urban area. *Remote Sensing of the Environment*, *61*, 361−370.

Lawrence, W. T., Imhoff, M. L., Kerle, N., & Stutzer, D. (2002). Quantifying urban land use and impact on soils in Egypt using diurnal satellite imagery of the Earth surface. *International Journal of Remote Sensing*, *23*, 3921−3937.

Martin, D. (1989). Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers*, *NS14*(1), 90−97.

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *Professional Geographer*, *55*, 31−42.

Nizeyimana, E. L., Petersen, G. W., Imhoff, M. L., Sinclair Jr., H. R., Waltman, S. W., Reed-Margetan, D. S., Levine, E. R., & Russo, J. M. (2001). Assessing the impact of land conversion to urban use on soils with different productivity levels in the USA. *Soil Science Society of America Journal*, *65*, 391−402.

Nunes de Lima, M. V. (Ed.). (2005). *IMAGE2000 and CLC2000. Products and methods. Corine land cover — updating for the year 2000. Ispra, European Commission, Joint Research Centre*.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem. CATMOG no. 38.* Norwich: Geo Books.

Openshaw, S., & Taylor, P. J. (1981). The Modifiable Areal Unit Problem. In N. Wrigley & R. J. Bennett (Eds.), *Quantitative geography; a British view* (pp. 60−70). London: Routledge and Kegan Paul.

Pozzi, F., Small, C., & Yetman, G. (2002). Modeling the distribution of human population with night-time satellite imagery and gridded population of the world. *Pecora 15/Land Satellite Information IV/ISPRS Commission I/ FIEOS 2002, Conference Proceedings*.

Small, C., Pozzi, F., & Elvidge, C. D. (2005). Spatial analysis of global urban extent from DMSP-OLS night lights. *Remote Sensing of Environment*, *96*, 277−291.

Sutton, P., Roberts, D., Elvidge, C., & Baugh, K. (2001). Census from heaven, an estimate of the global human population using night-time satellite imagery. *International Journal of Remote Sensing*, *22*, 3061−3076.

Tobler, W. R. (1979). Smooth pyncophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, *74*, 519−530.

Tobler, W. R., Deichmann, U., Gottsegen, J., & Maloy, K. (1995). *The Global Demography Project, Technical Report 95-6*. Santa Barbara: National Center for Geographic Information and Analysis.